# ANÁLISIS DE COMPONENTES PRINCIPALES Y DE CLÚSTER DE MUNICIPIOS PARA SERVICIOS FIJOS

## 1 Introducción - Capítulo metodológico

La prestación de los diferentes servicios fijos (Internet fijo, Telefonía fija y Televisión por suscripción) tiene condiciones heterogéneas a lo largo del territorio nacional. Los factores que pueden incidir en la prestación de estos servicios son ampliamente diversos, en el presente análisis los agregamos en 3 categorías: en primer lugar, pueden estar dados por el mismo comportamiento del mercado, como el número de empresas prestadoras del servicio o el número de hogares que lo demanden; en segundo lugar, por las condiciones socioeconómicas de los habitantes, las cuales pueden impactar en factores como la capacidad de pago, dinámicas de consumo, entre otras; por último, por las condiciones geográficas de la región o zona en donde ésta se evalúe, lo cual puede repercutir en las condiciones de despliegue de infraestructura, necesaria para garantizar la prestación del servicio.

Dado que gran parte de estos factores pueden ser observados a nivel municipal, el análisis simultáneo de dichos factores permite establecer grupos de municipios que sean lo más homogéneos posible (es decir, lo más similares a su grupo), pero a su vez lo más diferentes posible con otros grupos. Esta clasificación de municipios puede servir como insumo, no sólo para tener una caracterización clara de las condiciones de prestación de los servicios fijos, sino también como herramienta de focalización para programas de política pública.

Con el propósito de obtener una clasificación municipal como la mencionada, la CRC utiliza dos técnicas estadísticas multivariantes: un análisis de componentes principales (ACP), y un análisis de conglomerados o análisis de clúster. En este caso, la primera técnica se utiliza como insumo para el desarrollo de la segunda. Los aspectos más relevantes del desarrollo de estos dos tipos de análisis se exponen a continuación.

### 1.1 Análisis de Componentes Principales (ACP)

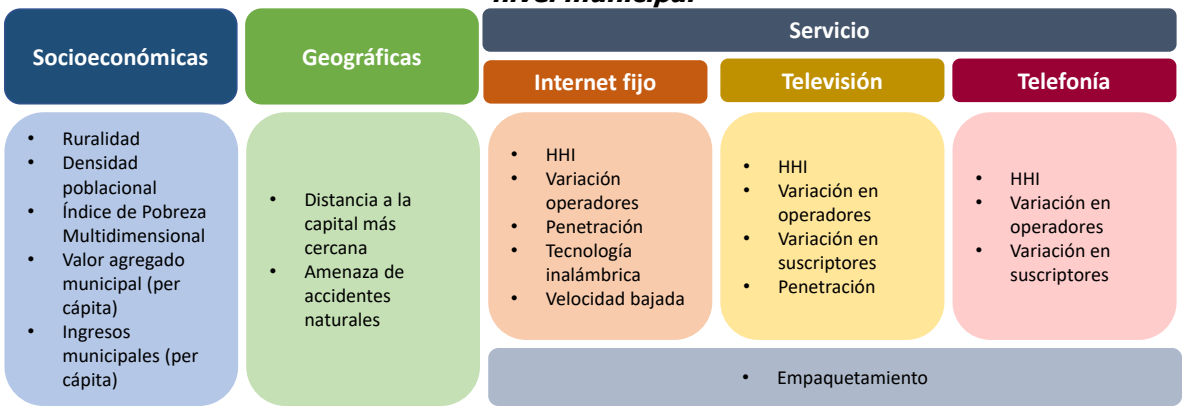
Dada la gran variedad de factores o variables que, como se mencionó previamente, pueden incidir en la prestación de los servicios fijos, puede ser útil, por una parte, determinar cuáles son aquellas que generan una mayor diferenciación entre los municipios y, por otra parte, y entendiendo la relevancia de las variables que se analizan, conservar la mayor cantidad de información que éstas proporcionan.

El análisis de componentes principales es una técnica estadística que permite tener un balance entre esos dos aspectos, de forma tal que se consolide la mayor cantidad de información posible que es proporcionada por las variables, identificando los aspectos que mejor capturan las diferencias entre los municipios. Esto se hace tomando todo el conjunto de variables de análisis y generando unas variables ficticias a partir de una combinación lineal de las variables iniciales; estas nuevas variables se conocen como Componentes Principales. Para entender mejor el procedimiento, en primer lugar, se realiza una descripción de las variables que se consideraron en el análisis y, posteriormente, se procede a explicar brevemente los aspectos más relevantes de la metodología utilizada.

### 1.1.1 Variables de estudio

Para desarrollar el presente estudio, se tuvieron en cuenta 23 variables, las cuales se pueden clasificar en 5 dimensiones: socioeconómica, geográfica, servicio de Internet fijo, servicio de televisión por suscripción y servicio de telefonía (ver Ilustración 1).

***Ilustración 1. Variables consideradas para la caracterización de los mercados fijos a nivel municipal***




**Fuente:** Elaboración CRC

## Variables socioeconómicas

- *Ruralidad:* Se refiere al porcentaje de población rural en el municipio, calculado con base en la información del Censo Nacional de Población y Vivienda (CNPV-2018) del DANE, y sus proyecciones poblacionales. Para capturar de forma estructural esta variable, se tomó el promedio de dicha participación, entre los años 2018-2021.
- *Densidad poblacional:* Medida como el número de habitantes del municipio, por kilómetro cuadrado, tomando como referencia el año 2021.
- *Índice de Pobreza Multidimensional (IPM):* Refleja la privación que tienen los individuos en cada municipio respecto a características consideradas como vitales, como salud, educación, empleo, entre otras[1]. En particular, se utilizó el IPM ajustado del DNP, el cual toma el Censo Nacional Agropecuario para el cálculo en zonas rurales, y el Censo Nacional para las cabeceras municipales[2].
- *Valor agregado municipal per cápita:* Se refiere a la diferencia entre la producción y el consumo intermedio del municipio[3], y sirve como proxy para establecer su comportamiento económico. Con el fin de minimizar los posibles efectos cíclicos y de atenuar el efecto de la cantidad de población, este valor fue convertido a precios constantes, se hizo el cálculo del valor agregado por habitante y se promedió el valor de los últimos cuatro años disponibles (2016-2019).

[1] COLOMBIA. DANE, Medida de pobreza multidimensional municipal de fuente censal. Boletín Técnico [en línea]. Bogotá, D.C., enero de 2020. Disponible en: <https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2018/informacion-censal/bt-censal-pobreza-municipal-2018.pdf>
[2] El DNP utiliza este índice como insumo en el cálculo del Índice Municipal de Riesgo.
[3] COLOMBIA. DANE. Ficha Metodológica: Cuentas Departamentales [en línea]. Bogotá, D.C., junio de 2016. Disponible en: <https://www.dane.gov.co/files/investigaciones/fichas/ficha_metodologica_CD-01_V5.pdf>

- *Ingresos municipales per cápita:* Se refiere a los ingresos totales del municipio (tributarios, no tributarios y transferencias). Los valores fueron obtenidos del Sistema de Información del Formulario Único Territorial (SISFUT) del DNP. Con el fin de minimizar los posibles efectos cíclicos y de atenuar el efecto de la cantidad de población, este valor fue convertido a precios constantes, se hizo el cálculo por habitante y se promedió el valor de los ingresos entre los años 2017-2020.

## Variables geográficas

- *Distancia a la capital más cercana:* Es la distancia carreteable, medida en kilómetros, que existe entre el municipio y la ciudad capital departamental más cercana[4]. Esta variable fue construida por la CRC a partir de la información de Google maps, consultada en noviembre de 2015.
- *Amenaza de accidentes naturales:* Se refiere a la proporción geográfica del municipio que es susceptible a inundaciones, movimientos en masa (por ejemplo, deslizamientos de tierra y derrumbes) y flujos torrenciales[5]. En este estudio se utiliza como proxy de las condiciones geográficas que pueden llegar a incidir en el acceso físico al municipio, y que a su vez pueden llegar a incidir en factores como el despliegue de infraestructura (tanto vial como del sector de telecomunicaciones).

## Variables de Servicios fijos (Internet fijo, telefonía fija y televisión por suscipción)

- *HHI:* Es una medida para determinar la concentración económica en cada uno de los servicios objeto de estudio. Esta variable se calcula a partir de las participaciones[6] de los operadores de cada uno de los servicios fijos.
- *Variación operadores:* Variación en niveles del número de operadores en cada uno de los servicios fijos objeto de estudio, entre los años 2017-2020.
- *Variación suscriptores*[7]*:* Variación en niveles del número de suscriptores de servicios en cada uno de los mercados objeto de estudio, entre los años 2017-2020.
- *Penetración*[8]*:* Se refiere al número de accesos en el mercado de Internet fijo, segmento residencial (o suscriptores en los otros mercados), dividido entre el número de hogares de cada municipio.
- *Empaquetamiento:* del total de servicios fijos suscritos en el año 2020 se calculó el porcentaje de servicios que se demanda en conjunto, es decir, planes tipo *duo play* o *triple play*.

[4] En el caso de los municipios en los que no fue posible encontrar una vía carreteable, y con el fin de capturar la dificultad del acceso a los mismos, esta variable fue imputada, calculando el valor máximo de las distancias encontradas y adicionando una desviación estándar.
[5] El DNP utiliza esta información como insumo en el cálculo del Índice Municipal de Riesgo. Utiliza como fuente información del Sistema Geológico Colombiano (SGC), del Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) y del Instituto Geográfico Agustín Codazzi (IGAC).
[6] Las participaciones se calculan a partir de las variables de accesos para el servicio de Internet fijo (2020), líneas para el servicio de telefonía (2020), y suscriptores para el servicio de televisión (2019).
[7] En el caso de Internet fijo, para el segmento residencial, se contempló introducir la variable del cambio de accesos. Sin embargo, esta variable estaba altamente correlacionada con otras variables, por lo que, para garantizar un mejor desempeño del modelo, se optó por omitirla del análisis.
[8] En el caso de penetración de telefonía, se encontró que esta variable estaba altamente correlacionada con otras variables, por lo que, para garantizar un mejor desempeño del modelo, se optó por omitirla del análisis.

Adicional a las variables anteriores, y específicamente para el caso del mercado de Internet fijo, se incluyeron las siguientes:

- *Tecnología inalámbrica:* Se refiere a la proporción de accesos en el municipio que corresponde a tecnologías inalámbricas (satelital, WiFi, WiMAX u otras) en el año 2020.
- *Velocidad de bajada*[9]*:* Se refiere a la velocidad de bajada promedio del municipio en el año 2020.

### 1.1.2 Componentes Principales

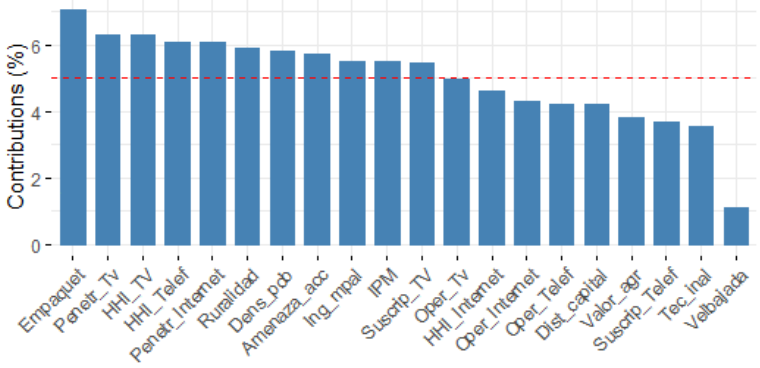
Como se mencionó previamente, al ACP es una metodología que permite consolidar toda la información proporcionada por las 23 variables consideradas, y "resumirlas" en una serie de variables ficticias, las cuales reciben el nombre de Componentes Principales (CP). Estas variables se construyen mediante una combinación lineal de las variables originales estandarizadas (ver Anexo 1.1).

La utilidad de esta "reducción de dimensión" de los datos radica principalmente en que permite extraer simultáneamente, de todas las variables analizadas, la información que captura en mayor proporción la variabilidad de los datos, es decir, la información que explica qué causa que los municipios sean diferentes entre sí, y capturarla o resumirla en un número inferior de variables. Dicho aspecto es útil para este estudio, pues en el desarrollo de modelos como los clústers, donde se suele recurrir a una selección de variables, la reducción que hace el ACP permite conservar la información más relevante del conjunto de variables observadas. Adicionalmente, por construcción, estas nuevas variables son independientes entre sí, propiedad estadística que permite una construcción más acertada de los clústers.

El ACP generalmente crea tantos CP como variables disponibles. A partir del análisis de los CP construidos con base en los datos municipales, el cual se explica brevemente en el Anexo 1.1, se optó por escoger los cinco primeros CP, los cuales capturan el 55,7% de la varianza de los datos. Estos cinco primeros componentes cuentan con información suministrada por todas las variables originales (ver Gráfico 1) [10].

[9] Se consideró incluir en el análisis la velocidad de subida y las desviaciones estándar de las velocidades en cada municipio. Sin embargo, estas variables estaban altamente correlacionadas con otras variables, por lo que, para garantizar un mejor desempeño del modelo, se optó por omitirlas.
[10] En el Gráfico 3 del Anexo 1.1.2 se muestra la contribución de las principales variables a cada uno de los CP de forma individual.

***Gráfico 1. Contribución de las variables a los primeros 5 CP en conjunto***




**Fuente:** Elaboración CRC

Con el fin de simplificar la estructura de los componentes (ver Anexo 1.1.2), se optó por hacer una rotación[11] de esos primeros cinco CP, los cuales, como se observa en la Tabla 1 invirtieron el grado de contribución de la varianza de los componentes originales, cambiando la varianza explicada de estos. Este cambio de varianza trae como resultado que el componente rotado 4 tenga una mayor contribución que el componente rotado 3. Al hacer una validación y al obtener un mejor desempeño en el ejercicio de clúster que prosigue a este análisis, se optó por conservar los componentes rotados 1, 2 y 4, los cuales en conjunto explican el 41,8% de la variabilidad de los datos, y que como se observa en la Tabla 8 del Anexo 1.1.2, explican la mayor parte de las variables utilizadas en el ejercicio[12].

***Tabla 1. Características de los primeros cinco componentes rotados***

| Atributo | RC1 | RC2 | RC4 | RC3 | RC5 |
|---|---|---|---|---|---|
| SS* cargas | 3.779 | 2.605 | 1.973 | 1.535 | 1.252 |
| Varianza explicada | 0,189 | 0,13 | 0,099 | 0,077 | 0,063 |
| Varianza acumulada explicada | 0,189 | 0,319 | 0,418 | 0,495 | 0,557 |

**Fuente:** Elaboración CRC
*SS: Suma de Cuadrados

## 1.2 Análisis de clúster

Ahora bien, con el objetivo de identificar los municipios que presentan características homogéneas o similares en la prestación de los servicios fijos, a nivel residencial, se aplicó la técnica estadística multivariante denominada Análisis de Conglomerados, también conocida como Análisis de Clúster.

El término *clustering* hace referencia a un amplio abanico de técnicas estadísticas multivariantes y *no supervisadas* cuya finalidad es encontrar patrones o grupos (*clusters*) dentro de un conjunto de observaciones. Las particiones se establecen de forma que las observaciones, en este caso los municipios, que están dentro de un mismo grupo, sean similares entre ellos y distintos a los municipios de otros grupos. En este sentido, los municipios que pertenecen a un conglomerado serán

[11] Se realizó una rotación ortogonal, mediante la metodología *varimax.*
[12] A excepción de la velocidad de bajada promedio municipal.

relativamente homogéneos entre sí, pero diferentes de aquellos que pertenecen a otros conglomerados. Esta técnica realiza la agrupación de tal forma que cada municipio pertenezca a un único grupo.

Al igual que el Análisis de Componentes Principales, el Análisis de Clúster es un método *no supervisado*, lo que implica que en el proceso para determinar la clasificación de las observaciones no se considera ninguna variable respuesta que indique de antemano la existencia previa de dicha clasificación (si es que existe tal variable). Por el contrario, las técnicas *supervisadas* emplean, para la determinación del modelo, un set de entrenamiento en el que se conoce de antemano la verdadera clasificación.

Dada la utilidad del *análisis de clúster* en diversas disciplinas (genómica, marketing, etc.), se han desarrollado multitud de variantes y adaptaciones de sus métodos y algoritmos, como se explica a continuación, en la actualidad, dichos métodos pueden agruparse principalmente en tres tipos[13]:

- *Partitioning Clustering*: Requiere que el usuario especifique de antemano el número de *clusters* que se van a crear (*K-means*, *K-medoids*, *CLARA*).
- *Hierarchical Clustering*: Este tipo de algoritmos no requiere que el usuario especifique de antemano el número de *clusters*. (*agglomerative clustering*, *divisive clusterig*).
- Métodos mixtos, los cuales combinan o modifican los anteriores (*hierarchical K-means*, *fuzzy clustering*, *model based clustering* y *density based clustering*).

Teniendo en cuenta lo anterior, y a partir de los tres componentes rotados que se obtuvieron a partir del ACP, se construyeron clústers mediante diferentes especificaciones (*k-means y k-medoids*, cada uno con diferente número de clústers, *jerárquicos aglomerativos*, y *jerárquicos* combinados con *k-means*). El desempeño de estos clústers fue analizado teniendo en cuenta tanto medidas de validación interna[14] como son los índices de conectividad, silueta y *Dunn*, como índices de estabilidad (*APN, AD, ADM, y FOM*)[15]; así mismo, se realizó un análisis de la caracterización de los clústeres, teniendo en cuenta las 23 variables consideradas al inicio de este ejercicio, ver Ilustración 1.

A partir de los análisis y validaciones mencionados, se determinó que el algoritmo de clustering que proporciona la mejor partición de los datos es un ***k-means* con 5 clústers**. El método K-means clustering encuentra los K mejores clústers, entendiendo como mejor clúster aquel cuya varianza interna (*intra-cluster variation*) sea lo más pequeña posible[16]. Se trata por lo tanto de un problema de optimización en el que se reparten las observaciones en K clústers, de forma tal que la suma de las varianzas internas de todos ellos sea lo menor posible.

[13] Amat, Joaquín (2017). Clustering y heatmaps: aprendizaje no supervisado. RPubs by RStudio. Disponible en: RPubs - Clustering y heatmaps: aprendizaje no supervisado con R

[14] Las técnicas de validación interna miden el clúster únicamente basadas en información de los datos y evalúan qué tan buena es la estructura del clúster. Estas medidas pueden usarse tanto para escoger el mejor algoritmo como para elegir el número de clúster óptimo.

[15] Ver Anexo 1.2

[16] MacQueen, 1967. Some methods for classification and analysis of multivariate observations. University of Califnornia, Los Angeles. Disponible en: https://books.google.com.co/books?hl=es&lr=&id=IC4Ku_7dBFUC&oi=fnd&pg=PA281&dq=MacQueen,+1967+k+means&ots=nPVcH_FanM&sig=xYiDSA9YWOv35j5dRAQ6ZROAd78#v=onepage&q=MacQueen%2C%201967%20k%20means&f=false

La Tabla 2 muestra la distribución tanto del número de municipios como de la población para los 5 clústers resultantes, los cuales se presentan ordenados en orden descendente de acuerdo con el desempeño de los municipios que los componen relativos a las condiciones a nivel socioeconómico, de acceso, y en la prestación de los servicios fijos, tomando como referencia cada una de las dimensiones y variables consideradas en el ejercicio de análisis de componentes principales (Ver anexo 1.1.2).

Así, se observa que el grupo de ***alto desempeño*** está conformado por 11 municipios, y aunque es el más pequeño de los clústeres al representar el 1% de los municipios, es a su vez el que agrupa la mayor proporción de población a nivel nacional (34,6%). El grupo de ***desempeño moderado*** está compuesto por 98 municipios, representa el 9% de los municipios, acumulando un 28,6% de la población a nivel nacional. Por otro lado, el grupo de ***desempeño incipiente*** está conformado por 219 municipios, un 20% del total de municipios de la muestra, que acumulan el 15,3% de la población nacional, mientras que el grupo de ***bajo desempeño***, que se consolida como el clúster más grande en términos geográficos al representar el 47% de los municipios del país, acumula tan solo un 14,8% de la población nacional. Por último, el grupo con un ***desempeño limitado*** está conformado por 268 municipios, representando un 24% del total de municipios, los cuales acumulan solo el 6,7% de la población.

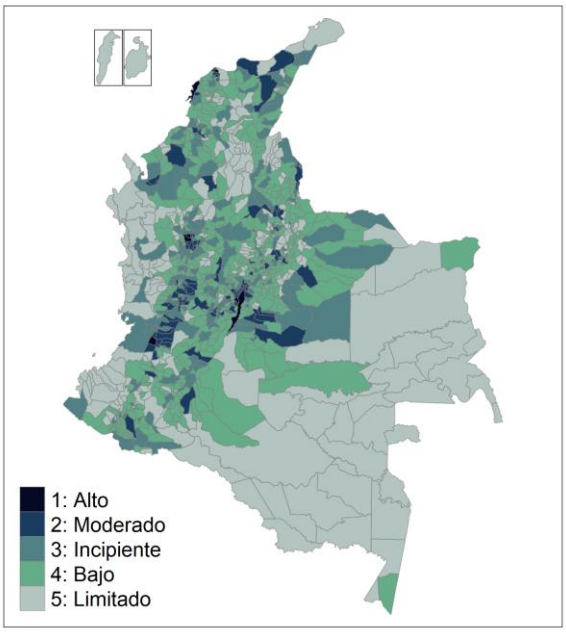
***Tabla 2. Tamaño de los grupos y porcentaje de población.***

| | Categoría por desempeño | Cantidad de municipios | Porcentaje de municipios | Población total | Porcentaje de población |
|---|---|---|---|---|---|
| **Modelos de K-medias** | Alto | 11 | 1% | 17.664.594 | 34,6% |
| | Moderado | 98 | 9% | 14.604.552 | 28,6% |
| | Incipiente | 219 | 20% | 7.833.950 | 15,3% |
| | Bajo | 526 | 47% | 7.532.868 | 14,8% |
| | Limitado | 268 | 24% | 3.413.534 | 6,7% |

**Fuente:** Elaboración CRC

La distribución geográfica de los clústers se observa en la Ilustración 2, donde se evidencia que los municipios con desempeño alto y moderado, se concentra alrededor de grandes ciudades y ciudades intermedias. Sin embargo, para entender mejor la caracterización de los clústers, a continuación, se realiza una breve descripción de cada uno.

***Ilustración 2. Clústers municipales de acuerdo con el desempeño en las dimensiones socioeconómicas, de acceso y de mercados fijos de servicios de telecomunicaciones***




**Fuente:** Elaboración CRC

**Municipios de alto desempeño:** Los municipios que fueron asignados a este grupo son ciudades capitales como Bogotá, Medellín, Barraquilla, Cartagena, y Cali; además municipios cercanos a estas capitales como Bello, Envigado, Itagüí, Sabaneta, Soledad, y Soacha. Este grupo se caracteriza por tener el menor porcentaje de ruralidad respecto de los demás grupos de municipios y una alta densidad poblacional. Adicionalmente, presentan en promedio un bajo porcentaje de población en condición de pobreza - 29% -, y un alto valor agregado per cápita; en promedio de 22,7 millones de pesos colombianos. Al considerar la variable de accesos, encontramos que la mayoría de los municipios tiene una baja proporción geográfica susceptible a inundaciones, movimientos en masa y flujos torrenciales -en promedio del 37%-. Por otro lado, son municipios con un nivel de penetración alto en los servicios fijos -Internet, 69%; Televisión, 61%; Telefonía 66%-, de igual forma, estos municipios presentan un grado de concentración entre el 0,39 y el 0,5. Respecto del total de servicios fijos que se demanda en estos municipios en promedio el 76% de estos consume de manera conjunta (duo play o triple play).

***Tabla 3. Estadísticas descriptivas de los municipios de alto desempeño.***

| Variables | Socioeconómicas | | | | Acceso | | Servicios | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ruralidad | Densidad poblacional | IPM | Valor agregado (Mill) | Amenaza | Distancia a la capital | Penetración TV | Penetración Telefonía | Penetración Internet | Empaquetados | HHI TV | HHI Telefonía | HHI Internet |
| **Media** | 0,04 | 6395,45 | 0,29 | 22,76 | 0,37 | 8,95 | 0,61 | 0,66 | 0,69 | 0,76 | 0,39 | 0,5 | 0,42 |
| **Desv. Estándar** | 0,05 | 3543,74 | 0,19 | 10,93 | 0,23 | 9,77 | 0,19 | 0,18 | 0,16 | 0,06 | 0,1 | 0,1 | 0,11 |

**Fuente:** Elaboración CRC

**Municipios con desempeño moderado:** Como se evidencia en la tabla 4, los municipios con desempeño moderado se caracterizan por tener un porcentaje de ruralidad moderado – en promedio del 19% - y una alta densidad poblacional. Adicionalmente, estos municipios presentan en promedio un bajo porcentaje de población en condición de pobreza – 30% -, también presentan un alto valor agregado per cápita; en promedio de 20 millones de pesos colombianos. Al considerar la variable de accesos, encontramos que la mayoría de los municipios son capitales[17] y cuidades intermedias, que tienen una baja proporción geográfica susceptible a inundaciones, movimientos en masa y flujos torrenciales -en promedio del 39%-. Por otro lado, estos municipios tienen un nivel de penetración medio en los servicios fijos, Internet fijo, 50%; Televisión por suscripción, 41%; Telefonía fija 34%, de igual forma, estos municipios presentan un grado de concentración entre el 0,38 y el 0,5. Respecto, del total de servicios fijos que se demanda en estos municipios en promedio el 66% de estos se consume de manera conjunta (duo play o triple play).

***Tabla 4. Estadísticas descriptivas de los municipios con desempeño moderado.***

| Variables | Socioeconómicas | | | | Acceso | | Servicios | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ruralidad | Densidad poblacional | IPM | Valor agregado (Mill) | Amenaza | Distancia a la capital | Penetración TV | Penetración Telefonía | Penetración Internet | Empaquetados | HHI TV | HHI Telefonía | HHI Internet |
| **Media** | 0,19 | 613,64 | 0,3 | 20,19 | 0,39 | 42,95 | 0,41 | 0,34 | 0,5 | 0,66 | 0,38 | 0,5 | 0,4 |
| **Desv. Estándar** | 0,15 | 708,2 | 0,15 | 16,13 | 0,23 | 44,09 | 0,15 | 0,17 | 0,15 | 0,11 | 0,1 | 0,2 | 0,13 |

**Fuente:** Elaboración CRC

**Municipios con desempeño incipiente:** Los municipios que componen este grupo se caracterizan por tener una proporción geográfica media susceptible a inundaciones, movimientos en masa y flujos torrenciales -en promedio del 44%-. A nivel de variables socioeconómicas, estos municipios presentan en promedio un porcentaje de ruralidad medio y una baja densidad población (Ver tabla 5). Además, estos municipios presentan una moderada proporción de la población con carencias; en promedio del 39%. De igual forma, son municipios con tasas de penetración menores a las del grupo de alto y

[17] Tunja, Manizales, Florencia, Popayán, Valledupar, Montería, Neiva, Riohacha, Santa Marta, Villavicencio, Pasto, Cúcuta, Armenia, Pereira, Bucaramanga, Sincelejo, Ibagué, y Yopal.

moderado desempeño -Internet fijo, 20%; Televisión por suscripción, 18%; Telefonía fija 7%- y el porcentaje de empaquetamiento es del 31% en promedio. Por último, estos municipios presentan un grado de concentración entre el 0,41 y el 0,90.

***Tabla 5. Estadísticas descriptivas de los municipios con desempeño incipiente.***

| Variables | Socioeconómicas | | | | Acceso | | Servicios | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ruralidad | Densidad poblacional | IPM | Valor agregado (Mill) | Amenaza | Distancia a la capital | Penetración TV | Penetración Telefonía | Penetración Internet | Empaquetados | HHI TV | HHI Telefonía | HHI Internet |
| **Media** | 0,41 | 111,12 | 0,39 | 19,08 | 0,44 | 82,36 | 0,18 | 0,07 | 0,2 | 0,31 | 0,41 | 0,9 | 0,52 |
| **Desv. Estándar** | 0,17 | 113,43 | 0,16 | 22,44 | 0,28 | 44,71 | 0,1 | 0,07 | 0,25 | 0,21 | 0,12 | 0,2 | 0,22 |

**Fuente:** Elaboración CRC

**Municipios con bajo desempeño:** Los municipios que fueron asignados a este grupo se caracterizan por tener un porcentaje de ruralidad alto -en promedio del 62%- y una baja densidad poblacional. Adicionalmente, estos municipios presentan un alto porcentaje de población en condición de pobreza; en promedio del 44%-, y presentan un moderado valor agregado per cápita; en promedio de 13,7 millones de pesos colombianos. Al considerar la variable de accesos, encontramos que en su mayoría son municipios alejados a las ciudades capitales, y tiene una alta proporción geográfica susceptible a inundaciones, movimientos en masa y flujos torrenciales en promedio del 54%-. Por otro lado, son municipios con un nivel de penetración bajo en los servicios fijos -Internet fijo, 5%; Televisión por suscripción, 10%; Telefonía fija, 1%-, de igual forma, estos municipios presentan un grado de concentración entre el 0,53 y 1, presentando una estructura de mercado cercana al monopolio en algunos casos.

***Tabla 6. Estadísticas descriptivas de los municipios de bajo desempeño.***

| Variables | Socioeconómicas | | | | Acceso | | Servicios | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ruralidad | Densidad poblacional | IPM | Valor agregado (Mill) | Amenaza | Distancia a la capital | Penetración TV | Penetración Telefonía | Penetración Internet | Empaquetados | HHI TV | HHI Telefonía | HHI Internet |
| **Media** | 0,62 | 55,92 | 0,44 | 13,79 | 0,54 | 97,8 | 0,1 | 0,01 | 0,05 | 0,04 | 0,53 | 1 | 0,63 |
| **Desv. Estándar** | 0,19 | 48,46 | 0,14 | 13,32 | 0,29 | 55,57 | 0,07 | 0,02 | 0,07 | 0,09 | 0,13 | 0 | 0,21 |

**Fuente:** Elaboración CRC

**Municipios con desempeño limitado:** La mayoría de los municipios que fueron asignados a este grupo se concentran en los departamentos de Vaupés, Guainía, Archipiélago de San Andrés, Providencia y Santa Catalina, Chocó, Vichada, Guaviare, Bolívar y Amazonas. Los municipios que componen este grupo se caracterizan y diferencian por ser aquellos que, en promedio, son los más alejados a las ciudades capitales, y tiene una alta proporción geográfica susceptible a inundaciones, movimientos en masa y flujos torrenciales -en promedio del 53%-. A nivel de variables

socioeconómicas, estos municipios presentan en promedio el más alto porcentaje de ruralidad -70%- y la menor cantidad de habitantes, respecto a los demás grupos. Además, en estos municipios se evidencia en promedio la mayor proporción de población con carencias -55%-, respecto de los demás grupos de municipios. De igual forma, son municipios con las menores tasas de penetración en los diferentes servicios fijos-Internet fijo, 3%; Televisión por suscripción, 7%; Telefonía 0%- y en los que existen únicamente uno o dos proveedores.

***Tabla 7. Estadísticas descriptivas de los municipios con desempeño limitado.***

| **Variables** | **Socioeconómicas** | | | | **Acceso** | | **Servicios** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ruralidad | Densidad poblacional | IPM | Valor agregado (Mill) | Amenaza | Distancia a la capital | Penetración TV | Penetración Telefonía | Penetración Internet | Empaquetados | HHI TV | HHI Telefonía | HHI Internet |
| **Media** | 0,7 | 39,51 | 0,55 | 9,96 | 0,53 | 215,21 | 0,07 | 0 | 0,03 | 0,03 | 0,87 | 1 | 0,81 |
| **Desv. Estándar** | 0,17 | 140,95 | 0,15 | 6,9 | 0,28 | 171,74 | 0,06 | 0,02 | 0,06 | 0,09 | 0,18 | 0 | 0,19 |

**Fuente:** Elaboración CRC

# ANEXOS

## 1.1. Aspectos metodológicos: Análisis de Componentes Principales (ACP)

El análisis de componentes principales es una técnica de aprendizaje no supervisado[18], en donde el conjunto de datos de interés se compone de una serie de variables que denotan características diferentes para una serie de individuos. Dadas estas variables, esta técnica es útil para[19]:

i. Comparar los individuos entre sí y detectar si existen patrones similares que permitan establecer subgrupos de individuos.
ii. Describir relaciones entre las variables, lo que permite entender cuáles son los patrones de similitud.
iii. Reducir la dimensión de representación. Esto se refiere a que, dada la relación que existe entre las variables, el ACP permite sintetizarlas en un número reducido de variables ficticias, que permitan resumir la información proporcionada por las variables originales. Este aspecto es útil para este estudio, pues en el desarrollo de modelos, como los clústers, donde se suele recurrir a una selección de variables, la reducción que hace el ACP permite conservar la información más relevante del conjunto de variables observadas.

Esas variables ficticias son conocidas como Componentes Principales (CP), y se construyen de forma tal que capturan la mayor varianza de las variables, es decir, la información que más contribuye a diferenciar a los individuos que se están analizando, que en este caso corresponde a los municipios del país.

### 1.1.1. Preparación de los datos y consideraciones previas

Partiendo del hecho de que el ACP se basa en capturar la varianza de los datos, es relevante mencionar que, dado que las variables originales tienen diferentes órdenes de magnitud[20] (ver Sección 1.1.1, en donde se describieron las variables empleadas para el estudio), éstas fueron estandarizadas de forma tal que tuvieran media cero y varianza uno. De no hacer este procedimiento, habría variables que podrían llegar a sobreestimar los CP, no porque realmente capturaran la varianza de los datos, sino por un efecto de su orden de magnitud.

Por otra parte, si existen variables con una correlación muy alta (mayor a 0,8), puede que estén brindando información muy similar lo cual puede llegar a sobredimensionar los CP. Por esta razón, se eliminaron variables que estaban inicialmente consideradas en el análisis, como lo son la velocidad de subida para el mercado de Internet fijo, las desviaciones estándar de dichas velocidades, la variación en niveles del número de accesos en los municipios, y la tasa de penetración del servicio de telefonía fija. Sin embargo, se resalta que buena parte de la información proporcionada por estas variables se está recogiendo de forma simultánea mediante otras que sí fueron incluidas.

[18] Se refiere a técnicas donde no existe una variable de interés independiente (o variable respuesta), a diferencia de métodos como la regresión lineal.
[19] PARDO, CAMPO ELÍAS. Estadística descriptiva multivariada. Bogotá, D.C.: Universidad Nacional de Colombia. Departamento de Estadística. Noviembre de 2015.
[20] Por ejemplo, unas variables están medidas en millones de pesos, otras en habitantes por km2, otras se construyeron como participaciones porcentuales, etc.

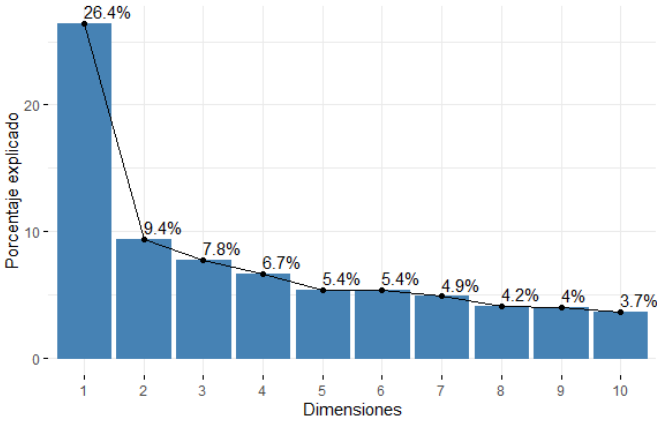
Finalmente se evaluó el test de *Kraiser, Meyer y Olkin (KMO)*, para determinar si la relación de las variables es apropiada para utilizar la metodología ACP. En conjunto, los datos tuvieron un *KMO=0,85*, e individualmente, las variables tuvieron un índice superior a *0,5*, indicando que los datos tienen una estructura apropiada para desarrollar este tipo de análisis[21].

### 1.1.2. Cálculo y selección de Componentes Principales

Los Componentes Principales se construyen mediante una combinación lineal de las variables originales, partiendo de fundamentos matemáticos de álgebra lineal. Así, los CP son vectores propios (o *eigenvectores*) que se toman de la matriz de correlaciones de las variables estandarizadas. Por construcción, el primer CP corresponde al vector que explica la máxima variabilidad de los datos; el segundo CP recoge la máxima variabilidad de los datos que no fue explicada por el primer CP; el tercer CP recoge la máxima variabilidad de los datos que no fue explicada por los primeros dos CP, y así sucesivamente[22]. Por lo general, existen tantos CP como número de variables originales, y la elección del número de CP que captura la mayor variabilidad de los datos, suficiente para el estudio, depende del investigador.

En este caso en particular, como se observa en el Gráfico 2, el primer CP captura el 26,4% de la varianza de los datos. De acuerdo con lo ya enunciado, la proporción de la varianza explicada por los siguientes CP va disminuyendo. Sin embargo, luego del quinto CP, la contribución de las siguientes dimensiones es marginal (lo cual se observa por el aplanamiento aparente de la curva que une las barras en el Gráfico). Adicionalmente, como se observa en la Tabla que acompaña al Gráfico, los valores propios asociados a esos primeros CP son mayores que 1, lo que indica que, en efecto, están brindando más información del que brindan las variables originales por separado. Así, con una varianza acumulada explicada del 55,7%, se decide tomar los primeros cinco CP para el resto del análisis.

***Gráfico 2. Variabilidad explicada por cada CP (%) – Primeros 10 CP***




| CP | Eigenvalor | Varianza explicada (%) | Varianza acumulada explicada (%) |
|---|---|---|---|
| Dim.1 | 5,29 | 26,44 | 26,44 |
| Dim.2 | 1,88 | 9,40 | 35,84 |
| Dim.3 | 1,56 | 7,78 | 43,62 |
| Dim.4 | 1,33 | 6,67 | 50,29 |
| Dim.5 | 1,09 | 5,44 | 55,72 |
| Dim.6 | 1,08 | 5,40 | 61,12 |
| Dim.7 | 0,98 | 4,91 | 66,03 |
| Dim.8 | 0,83 | 4,17 | 70,20 |
| Dim.9 | 0,80 | 4,02 | 74,22 |
| Dim.10 | 0,74 | 3,72 | 77,94 |

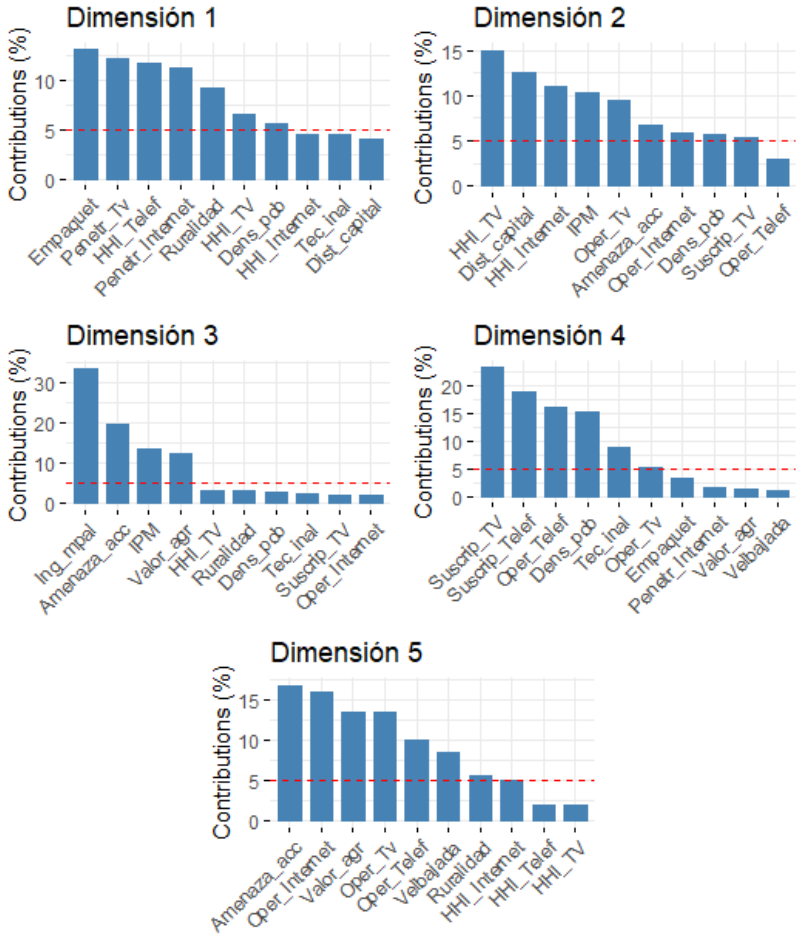
**Fuente:** Elaboración CRC

[21] REVELLE,W. psych: Procedures for Personality and Psychological Research [en línea]. Illinois: Universidad de Northwestern. < https://CRAN.R-project.org/package=psych> Versión=2.1.6.
[22] GIL M, CRISTINA. Análisis de Componentes Principales: RPubs by RStudio [en línea]. Junio de 2018. Disponible en <https://rpubs.com/Cristina_Gil/PCA>

Es relevante mencionar que la construcción de los CP se hace de forma tal que éstos son perpendiculares u ortogonales entre sí, lo que indica que no están correlacionados. Esta característica es importante, pues permite introducirlos en otros modelos, como en es el caso del análisis de clústers, minimizando el riesgo de una mala especificación.

El Gráfico 3 muestra las 10 variables que más contribuyen a cada uno de los cinco primeros CP. Así, se observa que las variables que más contribuyen al primer CP están relacionadas tanto con los tres mercados que se evalúan, como con variables socioeconómicas como la ruralidad del municipio y la densidad poblacional; el segundo CP se ve fuertemente influenciado tanto por las variables de concentración en el servicio de televisión por suscripción y en el mercado de Internet fijo, como por la variable de distancia a la capital más cercana; por su parte, el tercer CP se explica principalmente por variables socioeconómicas y geográficas, y así sucesivamente.

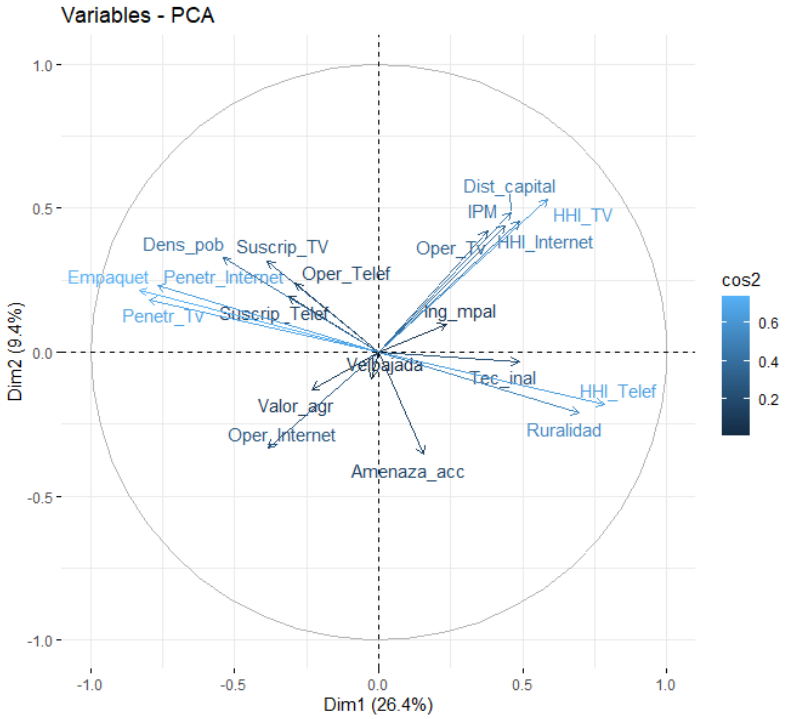
***Gráfico 3. Contribución de las 10 variables más importantes en cada uno de los 5 primeros CP***




**Fuente:** Elaboración CRC

El tipo de relaciones que captura el ACP está ejemplificado en el Gráfico 4, el cual muestra el círculo de correlación de los primeros dos componentes principales. En este Gráfico, las flechas más largas (dibujadas en color más claro) corresponden a las variables que tienen una mayor correlación con cada uno de los CP; haciendo énfasis en estas variables, se observa que la caracterización hecha a partir de estos dos primeros componentes indica que los municipios que tienen una mayor proporción de su población con programas de empaquetamiento de los tres servicios, junto con los municipios con mayor grado de penetración en los servicios de Internet fijo y Televisión por suscripción, tienden a tener mayor densidad poblacional que aquellos que no. Así mismo, este tipo de municipios tienen bajas tasas de ruralidad (lo que se evidencia porque la variable correspondiente se encuentra hacia el lado contrario de las primeras variables).

***Gráfico 4. Círculo de correlación CP1 vs CP2***




**Fuente:** Elaboración CRC

Este tipo de relaciones que se encuentran entre las variables son las que hacen que el ACP sea un buen insumo para, a partir de aquí, generar las diferentes agrupaciones en el análisis de clústers. Sin embargo, el análisis realizado, tanto a partir del Gráfico 3 como del Gráfico 4, no siempre resulta de fácil interpretación, principalmente debido al alto volumen de variables. En ese sentido, una práctica común en este tipo de análisis es la rotación de los factores[23] la cual permite tener una estructura simple, de forma tal que las variables originales se encuentren representadas en grupos (factores)

[23] Esta técnica, al ser utilizada como insumo para el análisis clúster que prosigue, brindó mejores resultados en términos de definición de los clústers.

mutuamente excluyentes, de modo que la contribución que realicen a esos componentes rotados sea alta en pocos de ellos, y baja en el resto[24].

En particular, en este caso se realizó una rotación ortogonal por *varimax*, la cual permite tener una representación como la mencionada en el párrafo anterior, de forma tal que los nuevos componentes tengan correlaciones altas con un pequeño número de variables, y correlaciones nulas en el resto, redistribuyendo la varianza de estos. La Tabla 8 muestra la matriz de cargas de los Componentes Rotados, de acuerdo con la cual se observa que la nueva varianza explicada por estos nuevos componentes se concentra en los RC (componentes rotados) 1, 2 y 4.

***Tabla 8. Matriz de cargas de los componentes rotados por varimax***

| Variable | RC1 | RC2 | RC4 | RC3 | RC5 |
|---|---|---|---|---|---|
| Oper_Tv | | 0,705 | -0,153 | 0,144 | |
| Oper_Telef | 0,587 | 0,227 | -0,188 | | 0,199 |
| Penetr_Tv | 0,703 | -0,243 | 0,374 | | |
| Empaquet | 0,823 | -0,216 | 0,2 | -0,145 | |
| HHI_Internet | -0,221 | 0,533 | | -0,163 | -0,383 |
| HHI_TV | -0,235 | 0,786 | | | -0,149 |
| HHI_Telef | -0,734 | 0,186 | -0,285 | | -0,144 |
| Suscrip_TV | 0,132 | | 0,762 | | |
| Suscrip_Telef | | | 0,629 | | |
| Ruralidad | -0,628 | 0,277 | -0,178 | 0,369 | 0,123 |
| Dens_pob | 0,29 | | 0,74 | | |
| Valor_agr | 0,174 | -0,271 | | 0,28 | -0,491 |
| Ing_mpal | -0,103 | 0,244 | | 0,544 | -0,492 |
| IPM | -0,288 | 0,459 | | -0,563 | |
| Dist_capital | -0,178 | 0,651 | | -0,103 | |
| Amenaza_acc | -0,152 | | | 0,759 | 0,189 |
| Penetr_Internet | 0,761 | -0,146 | 0,271 | | |
| Tec_inal | -0,572 | 0,144 | | -0,163 | 0,144 |
| Velbajada | | | | | 0,338 |
| Oper_Internet | 0,186 | -0,342 | | | 0,557 |

**Fuente:** Elaboración CRC

Tal como se enunció en la Sección 1.1.2, luego de hacer la rotación de los componentes, se optó por escoger estos tres componentes rotados, los cuales explican el 41,8% de la varianza de los datos. Este número de componentes fue el que tuvo un mejor desempeño en la construcción de los clústers municipales.

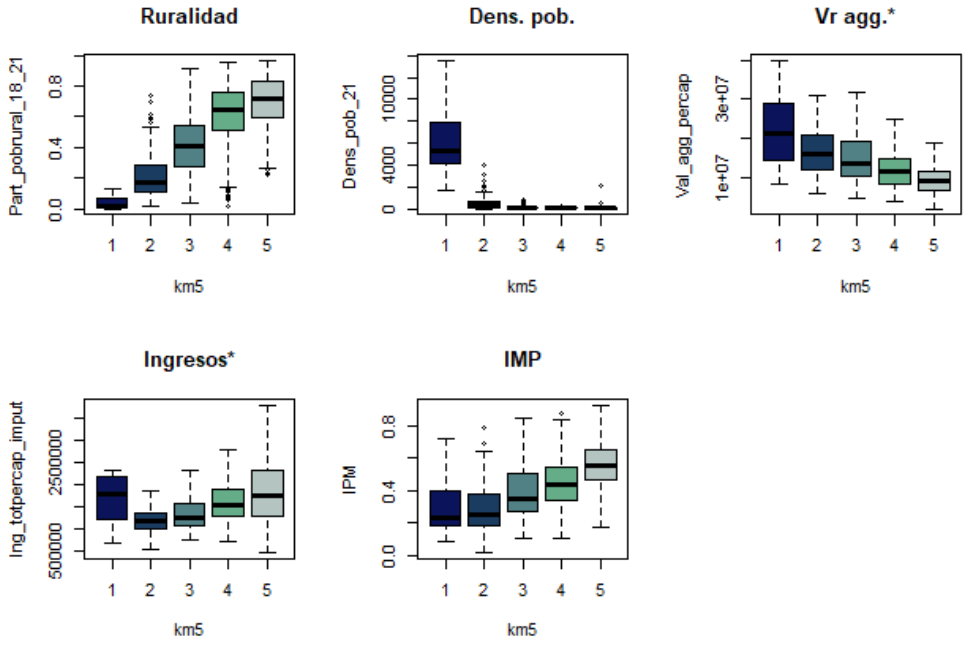
[24] MAVROU, IRINI. Análisis factorial exploratorio: cuestiones conceptuales y metodológicas. Revista Nebrija: Universidad Antonio de Nebrija [en línea]. 19. 2015. ISSN 1699-6569 Disponible en: <https://www.nebrija.com/revista-linguistica/analisis-factorial-exploratorio.html>

## 1.2. Evaluación de algoritmos a través de medidas de validación internas (conectividad, silhouette, Dunn y estabilidad)

| Algoritmo | Núm. de clúster | 3-clúster | 4-clúster | 5-clúster | 6-clúster | 7-clúster | 8-clúster | 9-clúster | 10-clúster |
|---|---|---|---|---|---|---|---|---|---|
| **Hierarchical** | APN | 0,0775 | 0,0566 | 0,0624 | 0,1651 | 0,167 | 0,1568 | 0,2781 | 0,2848 |
| | AD | 5,9567 | 4,7839 | 4,7438 | 4,6355 | 4,626 | 4,095 | 3,9065 | 3,7887 |
| | ADM | 0,8584 | 0,6935 | 0,7139 | 1,4274 | 1,4333 | 2,0426 | 2,0373 | 1,952 |
| | FOM | 2,902 | 2,4561 | 2,324 | 2,3021 | 2,2872 | 2,1372 | 2,0581 | 2,0325 |
| | Connectivity | 9,769 | 22,2841 | 25,3603 | 31,8385 | 34,7675 | 56,6726 | 60,8726 | 68,5496 |
| | Dunn | 0,1232 | 0,0357 | 0,0357 | 0,0427 | 0,0427 | 0,0216 | 0,0216 | 0,0279 |
| | Silhouette | 0,743 | 0,6332 | 0,6306 | 0,4148 | 0,3227 | 0,3671 | 0,3644 | 0,3507 |
| **K-medois (pam)** | APN | 0,2237 | 0,2582 | 0,3224 | 0,4101 | 0,4518 | 0,3643 | 0,4163 | 0,4098 |
| | AD | 3,9783 | 3,3718 | 3,2275 | 3,0548 | 2,9343 | 2,7418 | 2,6445 | 2,501 |
| | ADM | 1,5653 | 1,2221 | 1,4741 | 1,5062 | 1,5736 | 1,3696 | 1,3971 | 1,2584 |
| | FOM | 2,3001 | 2,017 | 1,9652 | 1,8763 | 1,827 | 1,802 | 1,7545 | 1,7346 |
| | Connectivity | 63,302 | 110,1206 | 163,2631 | 146,4202 | 166,6052 | 208,9901 | 221,025 | 228,4266 |
| | Dunn | 0,0035 | 0,0052 | 0,0028 | 0,0031 | 0,0031 | 0,0016 | 0,0023 | 0,0023 |
| | Silhouette | 0,4169 | 0,407 | 0,3222 | 0,3122 | 0,3325 | 0,337 | 0,3455 | 0,3382 |
| **K-means** | APN | 0,1714 | 0,3316 | 0,3124 | 0,3605 | 0,4095 | 0,3946 | 0,4087 | 0,3699 |
| | AD | 4,223 | 3,8432 | 3,4271 | 3,2715 | 3,1576 | 3,0138 | 2,9861 | 2,7209 |
| | ADM | 1,8893 | 2,0252 | 1,7657 | 1,7628 | 1,8587 | 1,6994 | 1,8513 | 1,4091 |
| | FOM | 2,2847 | 2,0554 | 2,0382 | 1,9279 | 1,8556 | 1,8044 | 1,8025 | 1,6121 |
| | Connectivity | 72,3667 | 90,9484 | 103,0353 | 109,1567 | 149,5218 | 139,5048 | 198,1198 | 189,8655 |
| | Dunn | 0,0035 | 0,0031 | 0,0033 | 0,0084 | 0,012 | 0,0082 | 0,0102 | 0,0142 |
| | Silhouette | 0,4053 | 0,4108 | 0,424 | 0,3981 | 0,3669 | 0,3501 | 0,3417 | 0,3366 |

## 1.3. Análisis de la composición de los grupos de municipios mediante boxplot

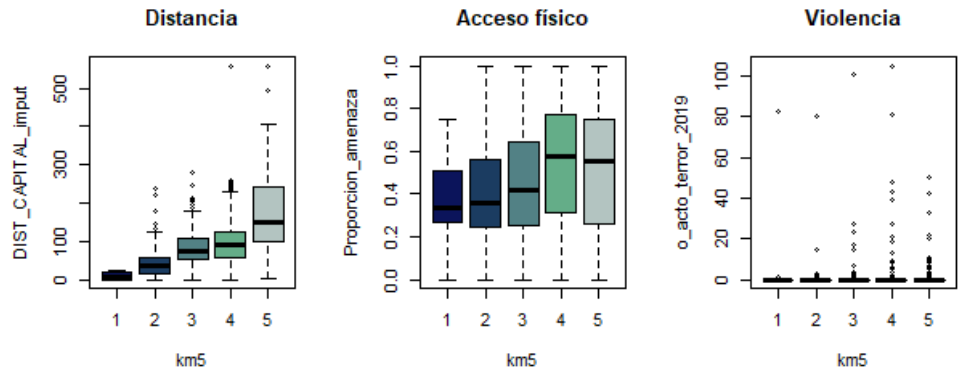
### 1.3.1. Variables socioeconómicas:




**Fuente:** Elaboración CRC

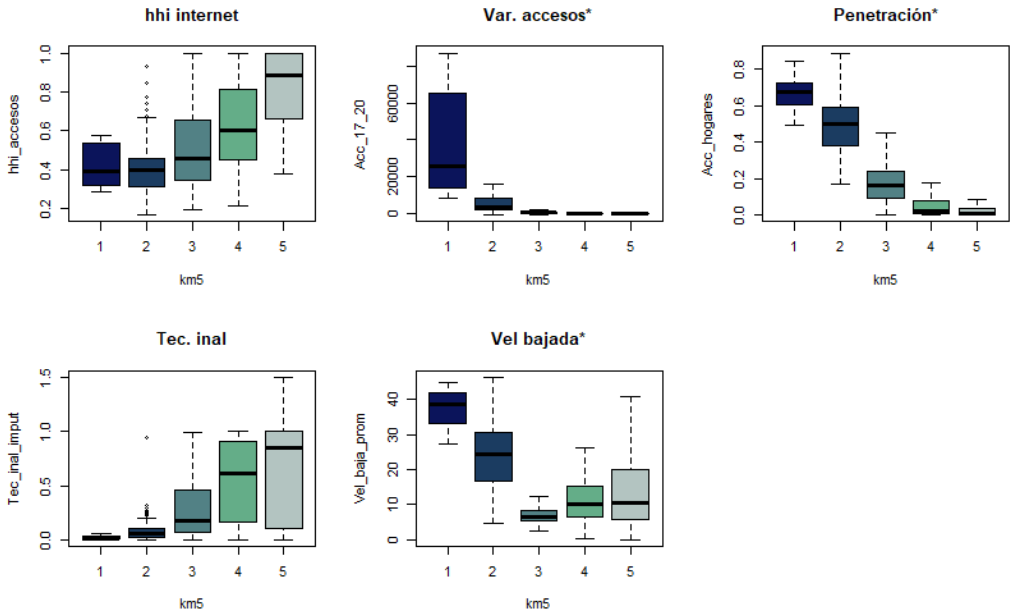
* Para mejor interpretación de los datos, las variables indicadas se graficaron omitiendo los valores atípicos.

### 1.3.2. Variables de acceso:




**Fuente:** Elaboración CRC

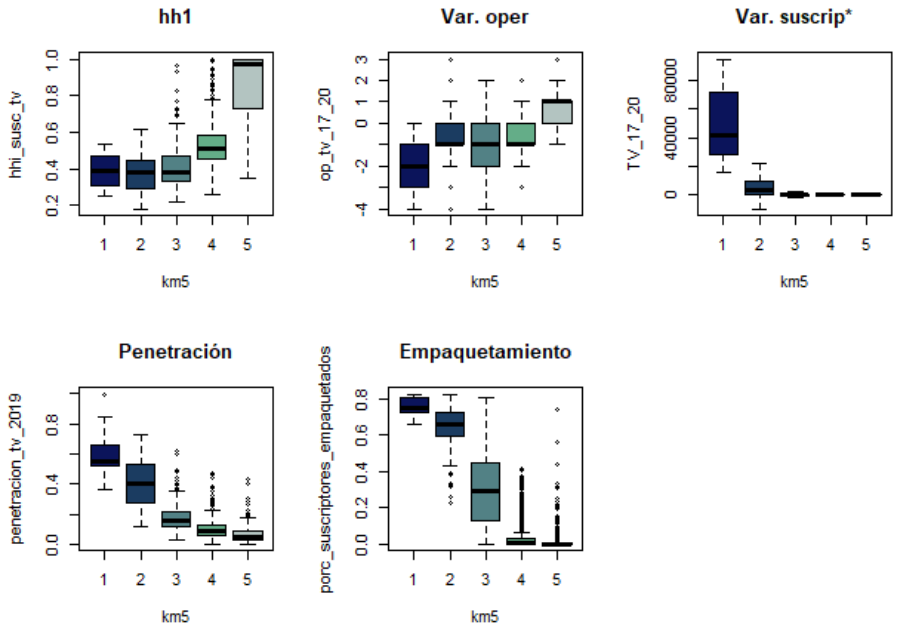
### 1.3.3. Variables servicio de Internet fijo:




**Fuente:** Elaboración CRC

* Para mejor interpretación de los datos, las variables indicadas se graficaron omitiendo los valores atípicos.

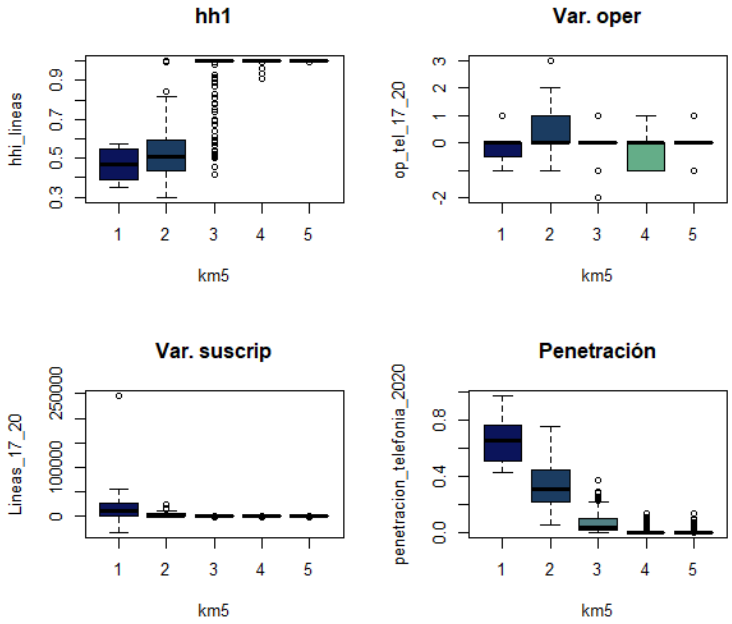
### 1.3.4. Variables servicio de Televisión:




**Fuente:** Elaboración CRC

* Para mejor interpretación de los datos, las variables indicadas se graficaron omitiendo los valores atípicos.

### 1.3.5. Variables servicio de Telefonía fija:




**Fuente:** Elaboración CRC